Open Problem: Regularized Optimization for Inducing Block-diagonal Matrices

Jin-ge Yao
Peking University
yaojinge@pku·edu·cn

Abstract

Many machine learning problems are parameterized by matrices with grouping effect. In the ideal cases, such matrices will naturally have a block-diagonal structure after permutation of rows and columns. However, real data are often produced with noise or perturbations. To learn robust parameter matrices, it is tempting to pose constraints or regularization terms in the optimization formulations to encourage for block-diagonal structures. The question posed here is whether there exists a generic way to induce block-diagonal structures for matrix-variable optimization problems.

1 Introduction

Many machine learning models or related problems are driven by parameter matrices with special grouping structures. *Spectral clustering* [1, 2, 3] and *subspace learning* [4, 5] aim at partitioning samples according to certain spectral properties of the affinity matrix. When the affinity matrix is close to block-diagonal, it has been shown that data points are tightly clustered in the eigenspace spanned by the first few eigenvectors of the Laplacian matrix. Algorithms for *kernel learning* problems [6, 7] try to learn pairwise similarities between data points, solving optimization problems for positive semidefinite matrices that typically addressed as kernel matrices. In *covariance selection* problems that involve learning *Gaussian graphical models* [8, 9] we need to find implicit conditional independence properties between covariates, resulting in a grouping structure in the covariance matrix or the precision matrix. In optimization problems with group-sparsity regularizers [10, 11], we need prior knowledge to divide features into groups before formulating the objective function. *Community detection* models [12, 13] try to discover latent groupings from a network or a graph represented as an affinity matrix.

For such problems, in the ideal cases where the underlying data can be well modeled without being collapsed by noise, the parameter matrix will have an exactly block-diagonal structure after permutation of rows and columns. Such block-diagonal parameterizations can benefit both computation and storage.

However, in real scenarios we do not have noiseless data. According to previous study [14], the main reason leading to the failure of spectral clustering on noisy data is that the block structure of the affinity matrix is destroyed by noise. This poses challenges to the clustering community: how to obtain correct clustering from noisy data, or further, how to obtain correct clustering from noisy data and remove the noise. Aiming at recovering the underlying block-diagonal structure, it is natural to ask whether we can find algorithms that are robust against noise, or furthermore, how to recognize the noise so as to remove them.

2 Open Problems

We assume that the related tasks can be parameterized by an affinity/similarity matrix A. Elements of A are typically pairwise similarities between data points or features/covariates. Alternatively, the parameters can be in the form of a self representation coefficient matrix $Z \in \mathbb{R}^{n \times n}$, which may be asymmetric in general. This setting exists in tasks such as subspace segmentation. In this case, a common way to form symmetric parametrization is to simply set $A = (|Z| + |Z^\top|)/2$.

Since the desired block-diagonal structures will appear only for a few particular orderings of rows and columns, in general we may not want to impose constraints directly over elements of the matrices. Instead, we may prefer to operate on their spectral properties. More specifically, we can control a matrix to be block-diagonal after permutation, via operating on its Laplacian matrix $L_A = D_A - A$, where D_A is a diagonal matrix of vertex degrees: $D_A = \text{diag}(\{d_i\}), d_i = \sum_j A_{ij}$.

There exists a well known result on the relationship between the number of blocks and the graph Laplacian of the matrix [15]:

Proposition. Let A be an affinity matrix. Then the multiplicity k of the eigenvalue 0 of the corresponding Laplacian L_A equals the number of connected components (i.e. blocks) in A.

Therefore, to induce block-diagonal structures, we only need to sparsify the eigenvalues of the Laplacian.

Define $\text{Lap}(A) := L_A$ as an operator that maps an affinity matrix A to its Laplacian.

Open Problem 1 (Regularization on Laplacian eigenvalues)**.** Is is possible to find a regularizer $\mathcal{R} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^+$ such that there exist algorithms that can efficiently find a local/global minimum of the optimization problem:

$$\min_A \quad \mathcal{L}(A) + \lambda \mathcal{R}(\text{Lap}(A)), \tag{1}$$

where $\mathcal{L}$ is a task-specific loss function that is possibly nonconvex or nonsmooth?

The first idea is to simultaneously minimize the rank of L_A as well as the loss function $\mathcal{L}$. Since exact rank minimization is in general NP-hard, we may choose $\mathcal{R}(A) = \|\text{Lap}(A)\|_*$, i.e. the trace norm of L_A. This is the convex surrogate for directly minimizing the rank of L_A. Natural coming-up problems are how can we solve such problems and under which conditions are we guaranteed to find the solutions.

Open Problem 2 (Sufficient conditions for guaranteed solutions)**.** How can we solve Problem 1 with the trace norm regularizer $\mathcal{R}(A) = \|\text{Lap}(A)\|_*$ and under which conditions on the loss function $\mathcal{L}$ are we guaranteed to find a local/global minimum efficiently?

3 Related Recent Progress

Currently related progress for inducing block-diagonal structures exists only for specific tasks, typically with the most commonly used regularizers. Some of the task-specific properties have been utilized, which make the conclusions not generalizable to the more general optimization scenarios.

3.1 Block-diagonal Covariance Selection

For the specific context of graphical lasso estimation, some independent findings [16, 17] have noticed that the block-diagonal structure of the precision matrix is determined by the block-diagonal structure of the thresholded empirical covariance matrix. For a given level of regularization λ, the original optimization problem (where S denotes the sample covariance matrix)

$$\min_{\Theta \succeq \mathbf{0}} \quad -\log\det(\Theta) + \text{tr}(S\Theta) + \lambda \|\Theta\|_{\ell_1} \tag{2}$$

can be solved with a simple procedure to induce block-diagonal solutions. First, the absolute values of the sample covariance matrix is thresholded at λ and set the thresholded matrix as an affinity matrix to detect strongly connected components. Then the original problem (2) is solved via separately maximizing the penalized log-likelihood on each component (block of covariates).

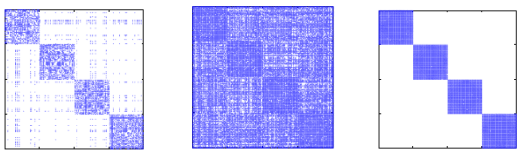

Figure 1: Grouping effects from different regularization settings. Image adapted from [19].

3.2 Subspace Clustering with Block-diagonal Affinity Matrices

Many related studies to enforce block-diagonal structures exist in the context of subspace segmentation. The task is formulated to pursue a sparse representation coefficient matrix Z given data matrix X by solving an optimization problem:

$$\min_{Z} \quad \|Z\|_{\alpha} + \frac{\lambda}{2}\|X - XZ\|_F^2, \text{ s.t. } \text{diag}(Z) = 0, \tag{3}$$

where $\alpha = 1$ or $*$ corresponds to sparse subspace clustering (SSC) [4] or low-rank representation (LRR) [5], respectively.

For sparse subspace clustering and low-rank representation models, it has been shown that the optimal solution matrix is block-diagonal under ideal conditions. Lu et al. [18] proposed the enforced block diagonal conditions, showing that any regularization term that satisfies these conditions will lead to a block-diagonal optimal solution. Unfortunately the analysis only holds for the most ideal cases where the data points are noiseless and subspaces are independent, while the open problem described in this paper is to guide the optimization process towards a block-diagonal solution given noisy or perturbed data.

Perhaps the most related attempt is the recent work [19] that tried to explicitly enforce block-diagonal structures. In that work, the authors imposed a fixed rank constraint on the graph Laplacian by restricting the solutions from what they called the k-block-diagonal-matrix set (k-BDMS):

$$\mathcal{K} = \{Z | \text{rank}(L_A) = n - k, A = (|Z| + |Z^\top|)/2\}. \tag{4}$$

In the above constraint set, the parameter k is the pre-specified number of subspaces. After building k block diagonal affinity matrix, the samples are readily segmented into k clusters.

The authors then employed the stochastic sub-gradient method to solve the optimization problems with the k-BDMS constraint. The key step is a projection onto k-BDMS, which was implemented with the augmented Lagrangian method (ALM) that leads to an iterative inner sub-procedure.

Figure 1 (adapted from [19] with permission) shows an illustrative comparison on synthetic noisy affinity matrices from 4 subspaces using different regularization settings: (a) ℓ_1 sparse regularization (SSC); (b) trace norm $\|\cdot\|_*$ (LRR); (c) block-diagonal constraint (k-BDMS).

The optimization problems involving k-BDMS are heavily nonconvex. Fortunately, it has been proved for both SSC and LRR that the simple stochastic sub-gradient method converges to the global optimum.

The most difficult part to generalize this methodology to other tasks is that the value of k needs to be specified before solving the optimization problem. We do not know the number of hidden clusters or groups in general settings. We cannot specify the level of sparsity of Laplacian eigenvalues, or equivalently the rank of Laplacian in advance [1].

Another issue is that the provable optimality guarantees are established only for specific cases: ℓ_2 loss for self representation and normal sparsity-inducing regularizers (ℓ_1 for SSC and nuclear norm for LRR). We don't know how these results will change in general settings from various tasks. Meanwhile, iteratively calling the ALM subprocedure to project current solution onto the k-block-diagonal-matrix set can make this optimization process difficult to be sufficiently scalable when applied to larger scale of data.

[1] This is analogous to sparsity-inducing regularization, where we typically do not have enough information to explicitly specify the number of non-zero elements in the final solution.

References

[1] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[2] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[3] Michael I. Jordan and Francis R. Bach. Learning spectral clustering. *Advances in Neural Information Processing Systems*, 16:305–312, 2004.

[4] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2765–2781, 2013.

[5] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010.

[6] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

[7] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. In *Advances in Neural Information Processing Systems*, pages 3149–3157, 2014.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[10] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[12] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

[13] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[14] Zhenguo Li, Jianzhuang Liu, Shifeng Chen, and Xiaoou Tang. Noise robust spectral clustering. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[15] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[16] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[17] Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, 13(1):781–794, 2012.

[18] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012*, pages 347–360. Springer, 2012.

[19] Jiashi Feng, Zhouchen Lin, Huan Xu, and Shuicheng Yan. Robust subspace segmentation with block-diagonal prior. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3818–3825. IEEE, 2014.